



Is your VLM Sky-Ready? A Comprehensive Spatial Intelligence Benchmark for UAV Navigation

Lingfeng Zhang^{1,2,3,*}, Yuchen Zhang^{3,4,*}, Hongsheng Li¹, Haoxiang Fu⁶, Yingbo Tang⁵
Hangjun Ye³, Long Chen³, Xiaojun Liang², Xiaoshuai Hao^{3,†,✉}, Wenbo Ding^{1,✉}

¹ Tsinghua Shenzhen International Graduate School, Tsinghua University

² Peng Cheng Laboratory ³ Xiaomi EV ⁴ Georgia Institute of Technology

⁵ Institute of Automation, CAS ⁶ National University of Singapore

zlf25@mails.tsinghua.edu.cn, haoxiaoshuai@xiaomi.com, ding.wenbo@sz.tsinghua.edu.cn

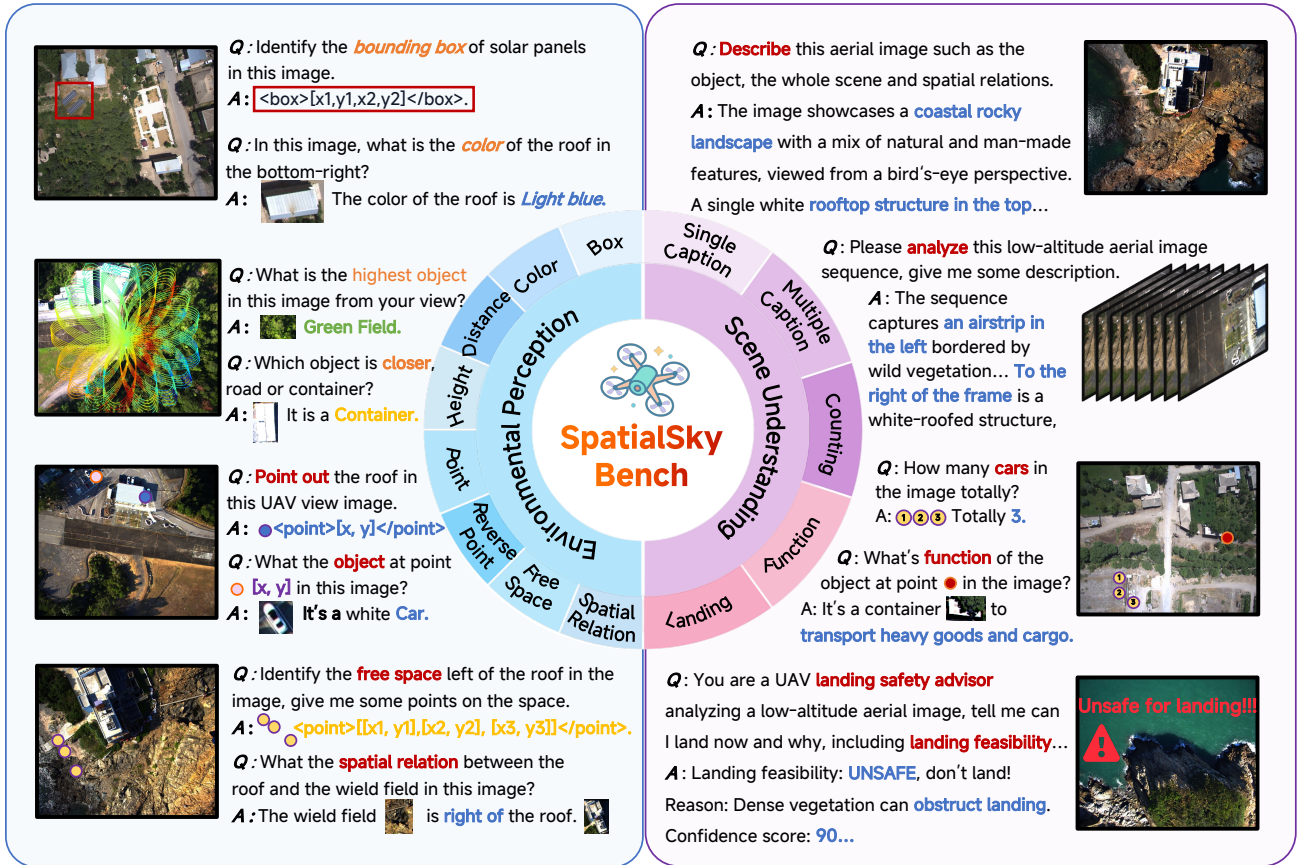


Figure 1. **Overview of SpatialSky-Bench**. Our benchmarks are divided into two categories: Environmental Perception and Scene Understanding, covering a total of 13 subcategories. We evaluated the VLM’s spatial intelligence capabilities across these UAV navigation tasks.

Abstract

Vision-Language Models (VLMs), leveraging their powerful visual perception and reasoning capabilities, have been widely applied in Unmanned Aerial Vehicle (UAV) tasks.

However, the spatial intelligence capabilities of existing VLMs in UAV scenarios remain largely unexplored, raising concerns about their effectiveness in navigating and interpreting dynamic environments. To bridge this gap, we introduce **SpatialSky-Bench**, a comprehensive benchmark specifically designed to evaluate the spatial intelligence capabilities of VLMs in UAV navigation. Our benchmark

* Equal contribution.

† Project leader.

✉ Corresponding authors.

comprises two categories—*Environmental Perception* and *Scene Understanding*—divided into 13 subcategories, including bounding boxes, color, distance, height, and landing safety analysis, among others. Extensive evaluations of various mainstream open-source and closed-source VLMs reveal unsatisfactory performance in complex UAV navigation scenarios, highlighting significant gaps in their spatial capabilities. To address this challenge, we developed the **SpatialSky-Dataset**, a comprehensive dataset containing 1 M samples with diverse annotations across various scenarios. Leveraging this dataset, we introduce **Sky-VLM**, a specialized VLM designed for UAV spatial reasoning across multiple granularities and contexts. Extensive experimental results demonstrate that **Sky-VLM** achieves state-of-the-art performance across all benchmark tasks, paving the way for the development of VLMs suitable for UAV scenarios. The source code is available at <https://github.com/linglingxiansen/SpatialSky>.

1. Introduction

Recently, the rapid development of Vision-Language Models (VLMs) has demonstrated their remarkable ability to understand and reason about visual scenes [1–15]. With the increasing prevalence of unmanned aerial vehicles (UAVs) in search and rescue operations, infrastructure inspection, and precision agriculture, VLMs have been successfully applied to UAV visual navigation tasks [16–27], showing promising application prospects. The spatial intelligence of VLMs is crucial for UAV navigation, enabling a detailed understanding of spatial relationships, fine-grained scene understanding, and precise environmental perception to support real-time UAV navigation decisions. However, existing VLM evaluation benchmarks primarily focus on human perspectives, such as indoor scenes, street scenes, and images taken with handheld cameras [11, 13, 28–36]. This difference in perspective makes existing benchmarks unable to assess the spatial intelligence capabilities of VLMs in UAV scenarios.

To bridge this gap, we propose **SpatialSky-Bench**, a comprehensive benchmark specifically designed to evaluate the spatial intelligence capabilities of VLMs in UAV navigation scenes. As shown in Fig. 1 and Fig. 2, our benchmark covers two main categories and thirteen fine-grained sub-capabilities, systematically evaluating the VLM’s understanding of UAV scenes. The first category is **environmental perception** capabilities, including (1) bounding box localization for accurate object detection; (2) target color recognition from a UAV perspective; (3) distance estimation between objects; (4) height perception from UAV view; (5) pointing to objects to locate the coordinates of a specific target; (6) pointing in reverse to identify objects at a given coordinate location; (7) free space detection to identify navigable areas; and (8) spatial relationship understand-

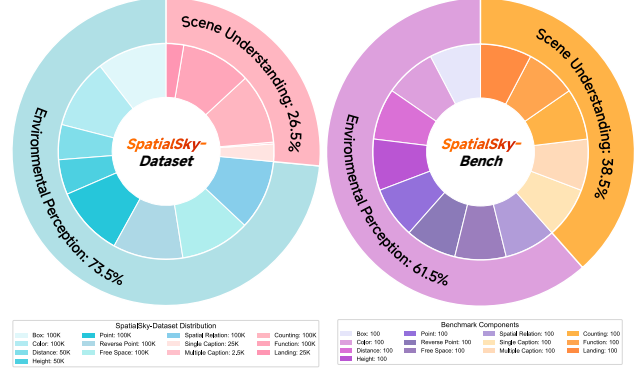


Figure 2. Distribution of our dataset and benchmark.

ing to determine the relative positions between targets. The second category is **scene understanding**, assessing the advanced cognitive abilities necessary for autonomous UAV navigation, including (9) scene captioning of a single UAV-view image, (10) time-series captioning of multiple images, (11) functional reasoning of objects in an image, (12) object counting at different scales and under occlusion, and (13) integrating overall spatial cues to determine whether a location is suitable for UAV landing.

To enhance the UAV spatial intelligence capabilities of VLM, we propose a scalable data generation method and introduce the **SpatialSky-Dataset**, a training dataset containing 1 M samples with diverse question-answer templates. Our data generation process utilizes multimodal inputs, including RGB images, semantic mask labels, LiDAR depth data, pose information, and bounding box annotations, to automatically generate origin dataset and question-answer pairs covering all 13 benchmark tasks. We then trained **Sky-VLM**, a spatial specific VLM for UAV navigation, using a two-stage training approach: first, we use supervised fine-tuning (SFT) on the **SpatialSky-Dataset** to acquire UAV-specific spatial reasoning capabilities; then, we add reinforcement fine-tuning (RFT) using Group Relative Policy Optimization (GRPO) [11, 12, 37, 38] to further optimize the model’s performance in key spatial reasoning tasks like box, pointing and counting. Extensive experiments demonstrate that **Sky-VLM** achieves state-of-the-art (SOTA) performance across all **SpatialSky-Bench** tasks, significantly outperforming existing open-source and closed-source VLMs in UAV scene spatial intelligence.

Our main contributions are summarized as follows:

- We propose **SpatialSky-Bench**, a comprehensive benchmark covering 2 main categories and 13 fine-grained sub-capabilities, for systematically evaluating the spatial intelligence of VLMs in UAV navigation scenarios.
- We construct **SpatialSky-Dataset**, a large-scale dataset containing 1 M samples generated through an automated process, including various annotation formats including open question-answer, multiple choice, pointing, and

bounding boxes, covering all benchmark tasks.

- We propose Sky-VLM, a dedicated UAV-view spatial awareness VLM trained using a two-stage approach: first, SFT to acquire UAV-specific spatial reasoning capabilities, and then GRPO to enhance its decision-making ability in complex navigation scenarios.
- Extensive experiments demonstrate that *Sky-VLM* achieves SOTA performance across all *SpatialSky-Bench* tasks, significantly outperforming both open-source and closed-source VLMs.

2. Related Work

VLM for UAV Navigation Unmanned aerial vehicle (UAV) navigation aims to enable UAVs to navigate autonomously based on high-level human commands. Traditional methods rely on supervised learning using human commands and flight trajectories collected in specific scenarios [39, 40]. Recently, VLM with its powerful visual language understanding capabilities has significantly advanced UAV navigation tasks [12, 25, 41]. For example, UAV-VLA [18] combines satellite imagery with the inference capabilities of VLMs to generate mission plans. See, Point, Fly, seepoint-fly [19] maps the output of VLMs to 3D waypoints through an intuitive visual pointing interface, achieving zero-shot UAV navigation. SoraNav [20] enriches the VLM input with geometric priors and switches between VLM inference and geometry-driven exploration based on navigation history; VLM-RRT [42] utilizes directions proposed by VLMs to guide RRT* sampling, accelerating path convergence. However, readily available VLMs generally lack accurate spatial awareness of UAV scenarios. For UAV navigation, spatial intelligence is crucial. It requires models to deeply understand spatial relationships, perform fine-grained scene analysis, and achieve accurate environmental perception, thereby supporting real-time flight decisions.

Spatial Intelligence Benchmark Recently, several spatial intelligence benchmarks have emerged to evaluate the spatial perception capabilities of VLMs for a wide range of tasks [3, 28–31, 43–52]. These benchmarks each have their own focus: early benchmarks like VQA [53] and GQA [54] emphasized semantic reasoning from static ground images. VSI-Bench [28] used indoor video to assess the changes in dynamic spatial memory over time. MMSI-Bench [29] tested spatial reasoning across multiple images using complex multi-step problems. RynnEC-Bench [30] extracted 22 region-based embodied cognition tasks from massive amounts of egocentric video. RefSpatial-Bench [31] introduced a spatial reference task involving up to five steps. RoboSpatial [3] provided a large-scale 2D/3D dataset with multi-perspective spatial annotations specifically for robotics-oriented spatial understanding, while Blink [43] highlighted that even basic perceptual capabilities underpinning such reasoning—like relative depth and visual cor-

respondence—remain poorly supported by current models. However, all existing spatial intelligence benchmarks share a common limitation: they all focus on spatial perception from a ground-based or egocentric perspective, neglecting the perception challenges from a UAV’s perspective. For example, there are challenges such as varying object scales, top-down occlusion, lack of depth information, and complex ground understanding requirements. To address this gap, we propose *SpatialSky-Bench* for evaluation and *SpatialSky-Dataset* for training.

3. SpatialSky Dataset and Benchmark Construction

3.1. Data Collection and Filtering

The *SpatialSky-Dataset* integrates annotations from UAVScenes [55], which include 20,000 images along with corresponding radar data and class labels at the mask level, covering 22 object categories. By combining 2D image masks with 3D LiDAR data, we create a diverse dataset for modeling object and spatial interactions. This dataset consists of two main components: *Environmental Perception* and *Scene Understanding*.

Environmental Perception Our environment-aware dataset incorporates eight fine-grained spatial reasoning capabilities, leveraging multimodal inputs including RGB images, semantic segmentation masks, LiDAR point clouds, and UAV pose information. For *bounding box localization*, we directly extract object instances from pixel-level semantic masks and transform each connected component into an axis-aligned bounding box (x_1, y_1, x_2, y_2) . For *color recognition*, we analyze the RGB distribution in each segmentation mask, calculating the dominant color by clustering pixel values in the HSV space and mapping them to descriptors like “light blue.” For pointing and reverse pointing tasks, we sample 5–8 pixel coordinates (x_i, y_i) within each mask region. The *reverse pointing* task uses these coordinates to identify the corresponding object category. For *free space detection and spatial relationship inference*, we utilize geometric properties of segmentation masks. We identify sufficiently large (>500 pixels) connected background regions and extract 3–5 points from each region. Spatial relationship inference is achieved by calculating the relative positions of object pairs. Given masks M_i and M_j , we compute their centroids $c_i = (\bar{x}_i, \bar{y}_i)$ and $c_j = (\bar{x}_j, \bar{y}_j)$, then determine the directional relationship using angle and distance thresholds:

$$\theta_{ij} = \arctan\left(\frac{\bar{y}_j - \bar{y}_i}{\bar{x}_j - \bar{x}_i}\right), \quad d_{ij} = \|c_i - c_j\|_2. \quad (1)$$

When d_{ij} exceeds a minimum threshold of 50 pixels, we categorize the relations into eight classes (left, right, top, etc.) and generate our original spatial relation dataset. For *distance estimation*, we directly utilize the LiDAR point

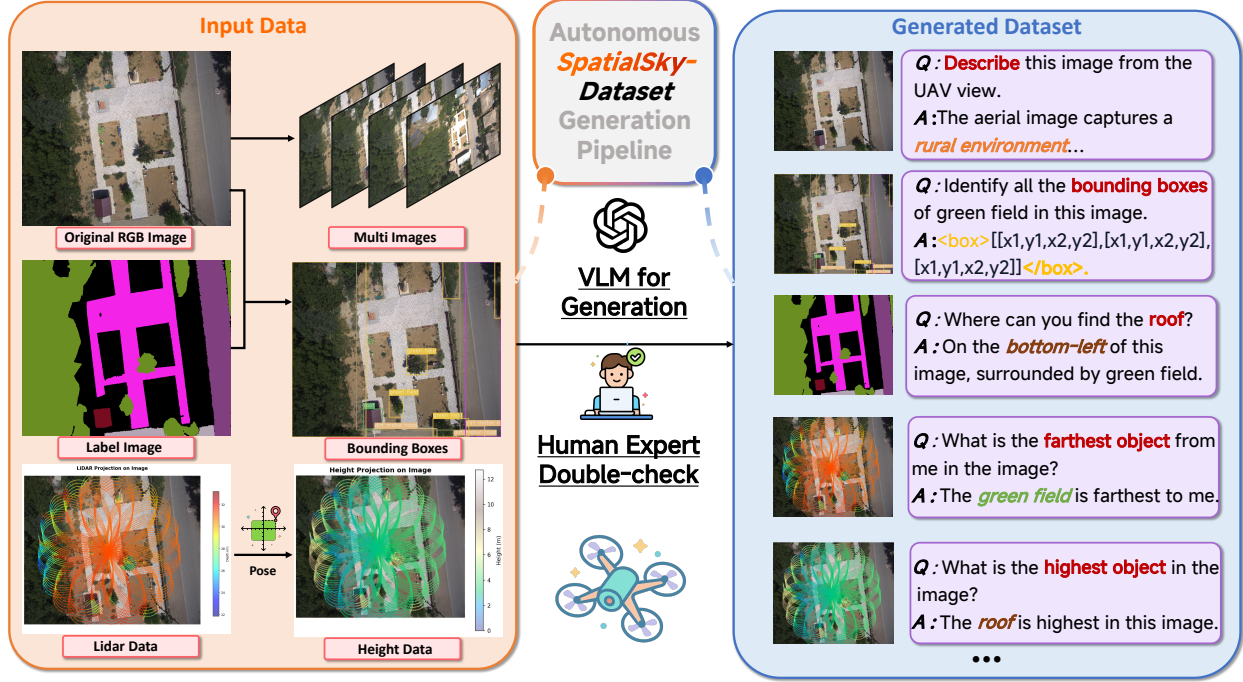


Figure 3. **SpatialSky-Dataset Generation Process.** Our generation pipeline take multimodal inputs, including RGB images, semantic labels, LiDAR depth data, UAV pose information, and bounding boxes. Using a VLM-based generation method and human expert validation, we automatically generate diverse question-answer pairs for 13 spatial reasoning tasks.

cloud: $\mathbf{P} = \{p_k\}_{k=1}^N$, where each point $p_k = (x_k, y_k, z_k)$ lies in the LiDAR coordinate system. We project the LiDAR points onto the image plane using the camera intrinsic parameters $\mathbf{K}$ and extrinsic parameters $[\mathbf{R}|\mathbf{t}]$, and then calculate the average depth:

$$d_{\text{obj}} = \frac{1}{|\mathcal{P}_{\text{obj}}|} \sum_{p_k \in \mathcal{P}_{\text{obj}}} z_k^{\text{cam}}, \quad (2)$$

where z_k^{cam} represents the depth value after converting the LiDAR points to camera coordinates. $p_k^{\text{cam}} = \mathbf{R} \cdot p_k$. To perform *height estimation*, we use the UAV pose transformation moments $\mathbf{T}_{4 \times 4}$ to convert the LiDAR point array into world coordinates. The global altitude of each point is obtained using the following formula:

$$\begin{bmatrix} x_k^w \\ y_k^w \\ z_k^w \\ 1 \end{bmatrix} = \mathbf{T}_{4 \times 4} \begin{bmatrix} x_k^{\text{cam}} \\ y_k^{\text{cam}} \\ z_k^{\text{cam}} \\ 1 \end{bmatrix}, \quad (3)$$

where z_k^w represents the absolute altitude in the world coordinate system. For each object, we calculate the average height to generate a height comparison query.

Scene Understanding Our scene understanding dataset comprises five high-level cognitive tasks that require holistic reasoning about aerial scenes, fully leveraging visual content and semantic context. For *single-image and multi-image captions*, we extract real-world object categories

from semantic masks and input single-image and multi-image sequences into a VLM, providing cues that emphasize aerial perspective features. The model generates captions covering scene composition, spatial object distribution, environmental context, and spatial variations within the multi-image sequence. For *object counting*, we apply connected component analysis to the semantic masks to identify individual instances $\{M_c^1, M_c^2, \dots, M_c^{n_c}\}$ for each category c . To ensure class balance, we employ stratified sampling, oversampling rare categories and undersampling major categories. For *function reasoning*, we manually craft 2–3 functional descriptions for each of the 22 object categories to reflect realistic drone scenes. We combine backpointing with functional queries by sampling points (x, y) within the object mask and generating questions. For landing safety analysis, we extract target distribution, available airspace ($>1,000$ pixels), potential hazards, and surface features, and then input them into the VLM. The model outputs a structured assessment, including feasibility classification (safe/cautious/unsafe), confidence score, recommended landing area, identified hazards and their risk levels, and comprehensive safety inference.

3.2. Question-Answer Pairs Generation

To ensure diversity and robustness of the benchmark, we designed task-specific QA formats for 13 fine-grained spatial reasoning tasks. We use VLM to generate more than 20 different question templates for each task, covering a vari-

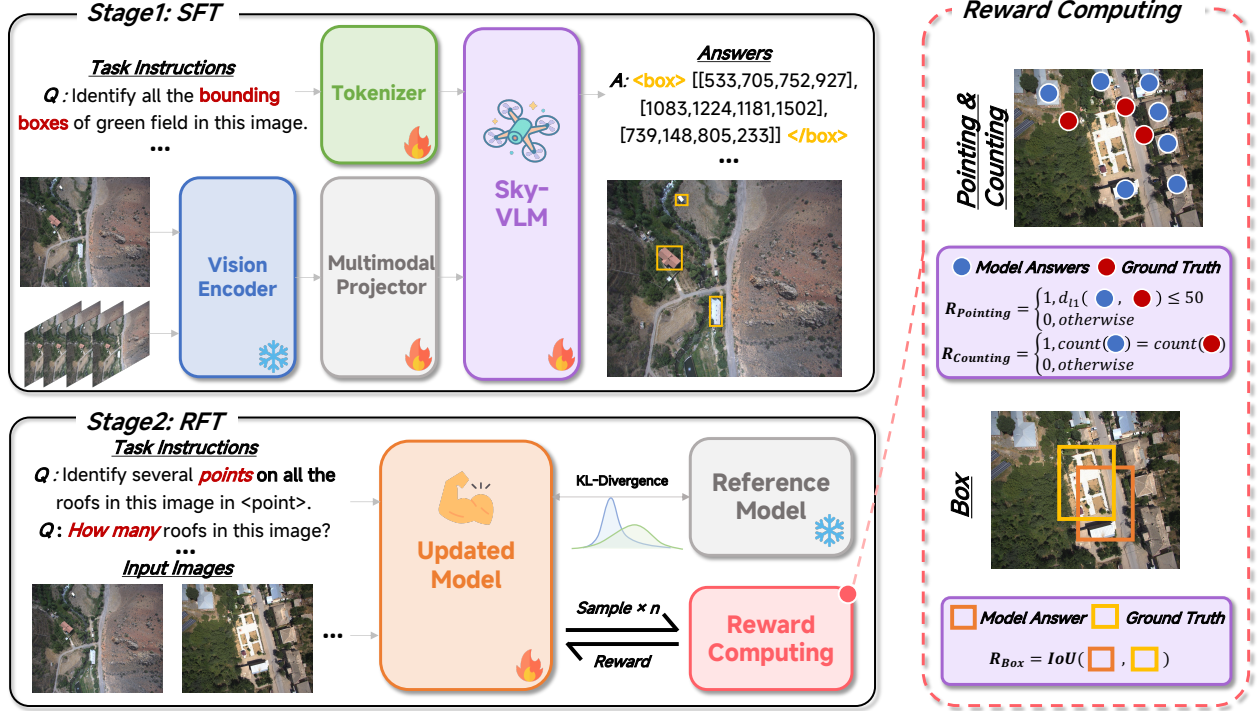


Figure 4. **Overview of our Sky-VLM.** Sky-VLM adopts a two-stage training approach. In the first stage, we involve supervised fine-tuning (SFT) on the entire **SpatialSky**-Dataset to develop the basic spatial reasoning capabilities. In the second stage, we use reinforcement fine-tuning (RFT), utilizing task-specific reward functions to enhance decision-making accuracy for key spatial tasks.

ety of linguistic expressions and query structures to prevent the model from relying on a single pattern match.

For the answer format, we adopted a structured representation to facilitate automatic evaluation. For bounding box and pointing tasks, we encapsulated coordinates in special tags: **<box>** $[[x_1, y_1, x_2, y_2], [x_1, y_1, x_2, y_2]]$ **</box>** and **<point>** $[[x_1, y_1], [x_2, y_2]]$ **</point>**. For color recognition, spatial relationships, and counting tasks, we constructed multiple-choice questions with 4–6 carefully crafted distractors and placed the answers within the **</boxed><choice>** tag. For distance estimation, height comparison, reverse pointing, scene caption, function, and landing safety tasks, we use open-ended questions with no strict restrictions, allowing for free responses. This multi-question design ensures that our benchmark comprehensively assesses structured spatial reasoning and open-ended semantic understanding capabilities.

3.3. SpatialSky-Bench

To construct a comprehensive and impartial evaluation benchmark, we carefully selected approximately 1,000 QA pairs from the generated dataset, ensuring balanced coverage of all 13 fine-grained tasks, 22 object categories, and various scene types. We employed stratified sampling to guarantee representativeness for each task category and scene context. Crucially, after selecting these benchmark samples, we removed all other QA pairs associated with the

same images from the training dataset, ensuring the benchmark consists entirely of unseen images to prevent data leakage and guarantee the fairness of the evaluation.

We designed task-specific evaluation metrics for each spatial reasoning ability. For bounding box localization, we calculate the Intersection over Union (IoU) between the predicted bounding box B_{pred} and the ground truth bounding box B_{gt} , and average it over all instances:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^N \frac{|B_{\text{pred}}^i \cap B_{\text{gt}}^i|}{|B_{\text{pred}}^i \cup B_{\text{gt}}^i|}. \quad (4)$$

If $\text{IoU} \geq 0.5$, the prediction is considered correct. For pointed tasks, we evaluate whether the predicted point falls within the true target mask M_{gt} . For multiple-choice questions and target object category recognition, we calculate the standard accuracy. For open-ended tasks, we use BLEU [56] and GPT-4o [57] as an automatic evaluator, providing it with the question, the true answer, and the model prediction, and then asking it to give a score from 1 to 10 based on factual correctness, semantic completeness, and reasoning quality. The final score is the average of all samples: $\text{Score}_{\text{open}} = \frac{1}{N} \sum_{i=1}^N s_i$, where $s_i \in [1, 10]$.

4. Sky-VLM

Framework We propose **Sky-VLM**, a VLM built on Qwen2.5-VL-7B [58] specifically designed for UAV spatial

reasoning tasks. The model employs a multimodal architecture flexibly handling both single-image and multi-image inputs for UAV spatial reasoning.

Supervised Fine-Tuning In the first phase of training, we performed supervised fine-tuning (SFT) on the entire *SpatialSky-Dataset* containing 1 million samples to establish the foundational spatial reasoning capabilities for UAVs. This phase enabled the model to: (1) learn aerial visual representations distinct from ground-based perspectives; (2) acquire task-specific output formats, including structured coordinates (box_i , point_i), multiple choice (boxed_i), and free descriptions; (3) develop basic spatial reasoning capabilities across all 13 benchmark tasks. We employed standard language modeling next-word prediction loss, but only computed gradients for the answer word to focus the learning on generating responses rather than understanding the question. Given a visual embedding sequence $\mathbf{V} = \{v_1, \dots, v_m\}$ and a text tag sequence $\mathbf{T} = \{t_1, \dots, t_n\}$, where the answer starts at position k , the SFT loss function can be expressed as:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{n-k+1} \sum_{i=k}^n \log P(t_i | \mathbf{V}, t_1, \dots, t_{i-1}; \theta), \quad (5)$$

where θ represents the model parameters, and $P(t_i | \cdot)$ represents the probability of predicting tag t_i given the visual context and preceding tags.

Reinforcement Fine-Tuning In the second stage, we apply Group Relative Policy Optimization (GRPO) to further improve the model’s decision-making ability and output accuracy in key spatial reasoning tasks. We constructed a reinforcement learning dataset containing 30,000 samples, focusing on tasks requiring precise localization and structured output. As shown in Fig. 4, we designed a task-specific reward function to directly measure the deviation between the model’s predictions and the ground truth labels.

For the pointing task, we evaluate the sequence of points output by the model, and the final task score is the average of the scores of all predicted points. The reward for a single point is binary, determined by calculating the L1 distance between the predicted point $(x_{\text{pred}}, y_{\text{pred}})$ and its nearest ground truth point (x, y) , using the following criteria:

$$R_{\text{point}} = \begin{cases} 1, & \text{if } |x_{\text{pred}} - x| + |y_{\text{pred}} - y| \leq 50, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

For multiple-choice tasks, we use an exact match reward:

$$R_{\text{choice}} = \begin{cases} 1, & \text{if predicted answer} = \text{true answer}, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

For bounding box localization, we calculate the IoU between the predicted and ground truth boxes as a continuous reward signal:

$$R_{\text{box}} = \text{IoU}(B_{\text{pred}}, B_{\text{gt}}) = \frac{|B_{\text{pred}} \cap B_{\text{gt}}|}{|B_{\text{pred}} \cup B_{\text{gt}}|}. \quad (8)$$

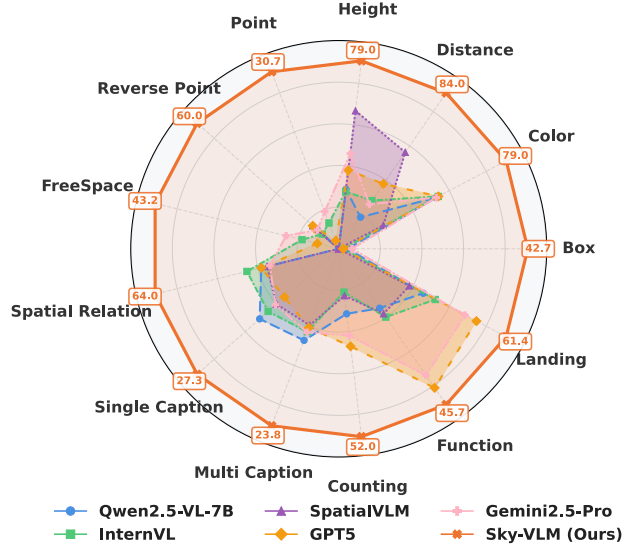


Figure 5. Performance of Our Sky-VLM.

The GRPO objective function aims to maximize the expected reward while maintaining closeness to the reference model π_{ref} through KL divergence regularization:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{\pi_{\theta}} \left[R(y) \cdot \log \frac{\pi_{\theta}(y|x)}{\pi_{\text{ref}}(y|x)} \right] + \beta \cdot \text{KL}(\pi_{\theta} || \pi_{\text{ref}}), \quad (9)$$

where β controls the strength of the KL penalty. Through the reinforcement learning phase, *Sky-VLM* is able to learn to generate more accurate spatial localization and consistent structured output, especially in tasks requiring pixel-level accuracy, where performance is significantly improved.

5. Experiments

5.1. Experimental Setup

Implementation Details Our *Sky-VLM* model is based on the Qwen2.5-VL-7B [58] and employs a two-stage process for initialization and training. The first stage involves supervised SFT on the *SpatialSky-Dataset* containing 1M samples. This stage utilizes eight H200 GPUs, the AdamW optimizer [65], a learning rate of 1e-5, a batch size of 2 per device, 2 gradient accumulation steps, and trains for one epoch. The second stage employs RFT using the GRPO [37] algorithm, trained on a dedicated dataset of 30K samples. In the RFT stage, a model with a learning rate of 1e-6 and weight decay of 0.1 is trained for one epoch, using the SFT model as the reference policy, with a KL regularization coefficient (β) of 0.01 to improve decision accuracy.

5.2. Evaluation Metrics

Our evaluation on *SpatialSky-Bench* employs a range of task-specific metrics for different spatial reasoning capabilities. For bounding boxes, we compute mIoU. For pointing

Model	Params	Environmental Perception								Scene Understanding					Avg.↑
		Box	Color	Dist.	Height	Point	Rev.	Free.	Sp. Rel.	Single	Multi	Cou.	Fun.	Land.	
Closed-source Models															
GPT-4-mini [57]	-	0.91	54.00	35.00	27.00	5.62	11.00	2.46	15.00	14.04	13.28	28.00	20.88	35.20	20.11
GPT-4o [57]	-	0.24	45.00	26.00	23.00	4.83	9.00	9.61	16.00	17.41	14.75	32.00	28.22	50.70	21.27
GPT-5 [57]	-	1.13	47.00	35.00	33.00	1.38	11.00	5.03	27.00	10.51	10.62	27.00	40.81	50.50	23.07
Gemini-2.5-Flash [59]	-	2.10	38.00	25.00	48.00	5.05	12.00	14.71	11.00	11.77	8.78	37.00	24.62	47.90	21.99
Gemini-2.5-Pro [59]	-	3.45	46.00	24.00	40.00	6.45	9.00	12.39	24.00	12.21	11.24	24.00	37.12	46.30	22.75
Qwen-VL-Max [58]	-	1.50	48.00	34.00	29.00	1.21	8.00	0.00	24.00	13.47	14.08	29.00	15.58	52.40	20.77
Open-source Models															
InternVL3.5 [60]	8B	1.41	47.00	26.00	24.00	4.48	7.00	8.64	32.00	13.64	11.19	12.00	20.05	35.40	18.65
Qwen3-VL-8B [61]	8B	1.32	12.00	26.00	44.00	6.46	3.00	6.93	33.00	11.07	12.30	12.00	26.16	4.33	15.25
Qwen2.5-VL-7B [58]	7B	2.38	46.00	17.00	25.00	0.79	9.00	0.00	27.00	15.27	12.33	18.00	17.43	31.32	16.93
Qwen2.5-VL-32B [58]	32B	0.27	47.00	16.00	13.00	9.88	6.00	2.34	21.00	10.78	6.98	0.00	32.67	15.50	13.93
Qwen2.5-VL-72B [58]	72B	0.23	53.00	23.00	25.00	11.08	6.00	3.25	22.00	12.22	10.08	0.00	33.08	10.11	16.05
Spatial Specific Models															
SpatialVLM [62]	8B	0.96	21.00	52.00	58.00	0.85	11.00	0.00	25.00	12.05	10.30	13.00	19.08	25.90	19.02
SpaceR [63]	7B	7.45	4.00	27.00	35.00	0.64	3.00	4.55	29.00	13.93	10.63	4.00	22.45	2.96	12.61
VILASR [64]	7B	2.08	0.00	37.00	42.00	3.92	7.00	3.87	26.00	12.92	12.92	0.00	24.33	2.76	13.45
Sky-VLM (Ours)	7B	42.68	79.00	84.00	79.00	30.72	60.00	43.20	64.00	27.34	23.83	52.00	45.72	61.40	53.30

Table 1. **Comparison Results of Various VLMs on SpatialSky-Bench.** Our *Sky-VLM* achieves SOTA performance. Dist., Rev., Free., Sp. Rel., Cou., Fun., Land., Avg., denote distance, reverse point, freespace, spatial relation, counting, function, landing and total average.

tasks, accuracy depends on whether the predicted coordinates lie within the mask of the real object. For tasks with discrete answers, such as multiple-choice questions and object category recognition, we report standard accuracy. Finally, for tasks such as image captioning and functional reasoning, we use BLEU-1 to BLEU-4 scores [56]. For landing, we use GPT-4o [57] as an automated evaluator. The final score for these tasks is the average of all sample scores.

5.3. Comparison with State-of-the-Art Models

We compared *Sky-VLM* with a comprehensive range of baseline models, including state-of-the-art closed-source models (GPT5 [57], Gemini2.5-Pro [59], *etc.*), open-source general-purpose VLMs (InternVL3.5 [60], Qwen2.5-VL [58], *etc.*), and spatial specific models (SpatialVLM [62], *etc.*). As shown in Tab. 1 and Fig 5, existing models perform poorly across all spatial inference tasks. The average scores of closed-source models range from 20.11 to 23.07, while open-source VLMs perform even worse (13.93 to 18.65). Even spatial specific models fail to effectively transfer to UAV perspectives; for example, SpatialVLM [62], SpaceR [63], and VILASR [64] achieve scores of only 19.02, 13.59, and 13.45, respectively. In comparison, our *Sky-VLM* model achieved SOTA performance across all the models, with an average score of **53.30**, **139.6%** improvement over the best baseline model (GPT-5, 23.07). *Sky-VLM* demonstrated superior performance across key tasks: bounding box score of 42.68 mIoU (**473%** improvement over SpaceR [63]), color score of 79.00 (**58%** improvement over SpatialVLM [62]), spatial relationship score of 70.00 (**38%** improvement over InternVL3.5 [60]), and landing score of 61.40 (**9%** improvement over Qwen-

Model	Params	Env. Per. Avg.↑	Sce. Und. Avg.↑	Total Avg.↑
<i>Other Models</i>				
Gemini-2.5-Pro [59]	-	20.61	26.17	22.75
Qwen2.5-VL-7B [58]	7B	15.75	18.81	16.93
SpaceR [63]	7B	13.75	10.84	12.78
<i>Our Models</i>				
Sky-VLM-SFT (Ours)	7B	52.53	41.52	48.29
Sky-VLM-RL (Ours)	7B	60.33	42.06	53.30

Table 2. **Ablation Study of Multi-Stage Training.**

VL-Max [58]). This demonstrates the powerful spatial intelligence capabilities of our model in UAV scenarios.

5.4. Ablation Study

Effect of Multi-Stage Training To verify the effectiveness of our proposed two-stage training method, we compared Sky-VLM-SFT, trained with only SFT, with Sky-VLM-RL, which incorporates reinforcement learning. As shown in Tab. 2, adding GRPO-based RFT significantly improved performance on spatial reasoning tasks (**53.30** vs. 48.29). Sky-VLM-RL achieved 60.33 score on the environment perception task, a 14.8% improvement over Sky-VLM-SFT (52.53), while maintaining similar performance on the scene understanding task (42.06 vs. 41.52).

Effect of Reward Model To demonstrate the role of each reward function in GRPO training, we conducted ablation experiments by removing individual reward components. As shown in Tab. 3, all three reward functions play a crucial role in achieving optimal performance. Removing the point reward resulted in the most significant perfor-

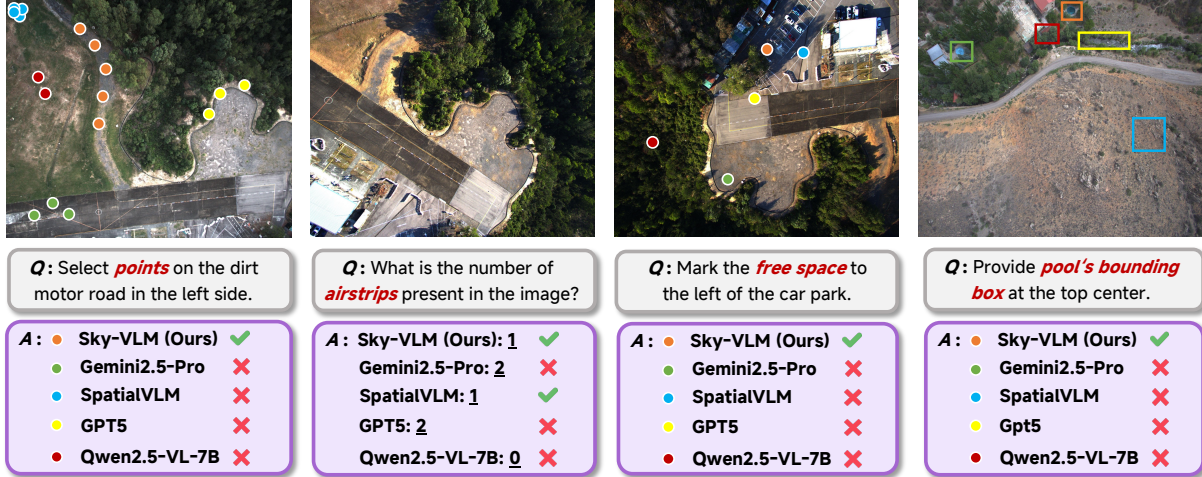


Figure 6. Qualitative Results of Different VLMs on *SpatialSky-Bench*.

Model	Env. Per.	Scs. Und.	Total
	Avg.↑	Avg.↑	Avg.↑
Qwen2.5-VL-7B [58]	15.75	18.81	16.93
Sky-VLM-SFT	52.53	41.52	48.29
Sky-VLM-RL w/o Box Reward	57.66	41.78	49.72
Sky-VLM-RL w/o Point Reward	53.77	40.95	47.36
Sky-VLM-RL w/o Multi-Choice Reward	59.32	41.22	50.27
Sky-VLM-RL (Ours)	60.33	42.06	53.30

Table 3. Ablation Study of Reward Model.

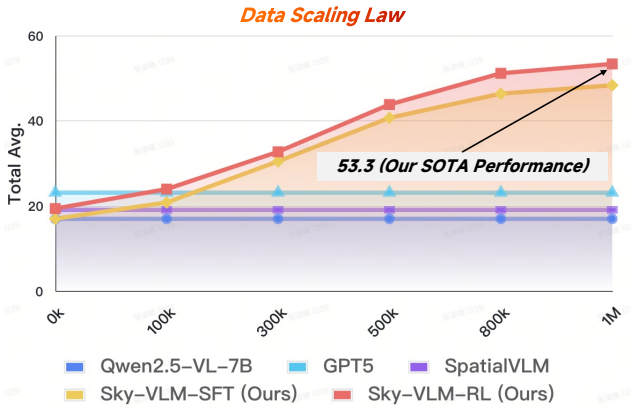


Figure 7. Data Scaling Law.

mance degradation, with the environment perception score dropping from 60.33 to 53.77 (6.56% improvement), indicating that accurate coordinate prediction is fundamental to spatial reasoning. The bounding box reward and multiple-choice reward are equally important; removing them reduced the total average to 49.72 and 50.27, respectively, demonstrating their criticality for accurate object localization and discrete decision-making tasks such as color recognition and spatial relationships.

Data Scaling Law To investigate the impact of training data size on model performance, we compared datasets with different sample sizes from 0K to 1M. As shown in Fig. 7, our model performance improved rapidly in the initial stage, then gradually saturated. Sky-VLM-SFT showed a signifi-

cant improvement in the early stages, with accuracy jumping from 16.93 (baseline) to 30.43 using only 300K samples. Accuracy gradually decreased with increasing sample size, reaching 48.29 with 1 million samples. More importantly, the reinforcement learning stage consistently improved SFT performance across all data sizes. With 100K samples, Sky-VLM-RL achieved a score of 23.9, while SFT achieved 20.77 (3.13% improvement); on the complete 1 million sample dataset, accuracy reached **53.3**, while SFT achieved 48.29 (**5.01%** improvement). This continuous improvement gap demonstrates that RFT can effectively enhance spatial reasoning capabilities.

5.5. Qualitative Analysis

Fig. 6 presents a qualitative comparison of four representative spatial reasoning tasks, demonstrating that *Sky-VLM* outperforms baseline models. In pointing task, Sky-VLM accurately identifies multiple valid locations on the road surface, while other models fail to provide correct predictions. In counting task, Sky-VLM correctly identifies one runway, while other models provide inaccurate counts, highlighting the challenge of object recognition from an aerial perspective. In freespace task, Sky-VLM successfully locates an open area, while all baseline models fail to complete this task. Furthermore, in box task, Sky-VLM generates an accurate bounding box at the top center, while baseline models are completely unable to detect or locate the target object. These visualizations clearly demonstrate that Sky-VLM’s spatial intelligence capabilities in UAV scenarios are significantly superior to other models.

6. Conclusion

We propose *SpatialSky-Bench*, which covers 13 fine-grained spatial reasoning tasks, categorized into environmental perception and scene understanding. Our extensive evaluation of mainstream VLMs reveals their signifi-

cant limitations in spatial intelligence when handling UAV perspectives, highlighting the unique challenges posed by UAV navigation scenarios. To address these challenges, we developed the *SpatialSky-Dataset* dataset, containing 1 million automatically generated samples with diverse annotation methods, and propose *Sky-VLM*. *Sky-VLM* is a specific VLM trained using a two-stage approach, combining SFT and RFT, supplemented by task-specific rewards. Extensive experimental results demonstrate that *Sky-VLM* achieves state-of-the-art performance across all benchmark tasks, significantly outperforming both open-source and closed-source VLMs in UAV spatial reasoning, paving the way for developing spatial-aware VLMs in UAV scenarios.

References

- [1] Yingbo Tang, Lingfeng Zhang, Shuyi Zhang, Yinyao Zhao, and Xiaoshuai Hao. Roboafford: A dataset and benchmark for enhancing object and spatial affordance learning in robot manipulation. In *Proceedings of ACM International Conference on Multimedia*, pages 12706–12713, 2025. 2
- [2] Yujie Wu, Huaihai Lyu, Yingbo Tang, Lingfeng Zhang, Zhihui Zhang, Wei Zhou, and Siqi Hao. Evaluating gpt-4o’s embodied intelligence: A comprehensive empirical study. *TechRxiv preprint techrxiv.174495686.69962588/v1*, 2025.
- [3] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. In *RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics*, 2025. 3
- [4] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1724–1734, 2025.
- [5] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.
- [6] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- [7] Yingbo Tang, Shuaike Zhang, Xiaoshuai Hao, Pengwei Wang, Jianlong Wu, Zhongyuan Wang, and Shanghang Zhang. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. *arXiv preprint arXiv:2503.00778*, 2025.
- [8] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Listen and look: Multi-modal aggregation and co-attention network for video-audio retrieval. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2022.
- [9] Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, Weiping Wang, and Dan Meng. What matters: Attentive and relational feature aggregation network for video-text retrieval. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2021.
- [10] Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*, 2025.
- [11] Yujie Wu, Huaihai Lyu, Yingbo Tang, Lingfeng Zhang, Zhihui Zhang, Wei Zhou, and Siqi Hao. Evaluating gpt-4o’s embodied intelligence: A comprehensive empirical study. *Authorea Preprints*, 2025. 2
- [12] BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 2, 3
- [13] Shuyi Zhang, Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Pengwei Wang, Zhongyuan Wang, Hongxuan Ma, and Shanghang Zhang. Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12745–12752, 2025. 2
- [14] Hao Cheng, Erjia Xiao, Yichi Wang, Lingfeng Zhang, Qiang Zhang, Jiahang Cao, Kaidi Xu, Mengshu Sun, Xiaoshuai Hao, Jindong Gu, et al. Exploring typographic visual prompts injection threats in cross-modality generation models. *arXiv preprint arXiv:2503.11519*, 2025.
- [15] Qiang Zhang, Zhang Zhang, Wei Cui, Jingkai Sun, Jiahang Cao, Yijie Guo, Gang Han, Wen Zhao, Jiaxu Wang, Chenghao Sun, et al. Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots. *arXiv preprint arXiv:2503.09010*, 2025. 2
- [16] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yan-ning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15384–15394, 2023. 2
- [17] Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. In *The International Conference on Learning Representations*, 2025.
- [18] Oleg Sautenkov, Yasheerah Yaqoot, Artem Lykov, Muhammad Ahsan Mustafa, Grik Tadevosyan, Aibek Akhmetkazy, Miguel Altamirano Cabrera, Mikhail Martynov, Sausar Karaf, and Dzmitry Tsetserukou. Uav-vla: Vision-language-action system for large scale aerial mission generation. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 1588–1592. IEEE, 2025. 3
- [19] Chih Yao Hu, Yang-Sen Lin, Yuna Lee, Chih-Hai Su, Jie-Ying Lee, Shr-Ruei Tsai, Chin-Yang Lin, Kuan-Wen Chen, Tsung-Wei Ke, and Yu-Lun Liu. See, point, fly: A learning-free vlm framework for universal unmanned aerial navigation. In *Conference on Robot Learning*, pages 4697–4708, 2025. 3
- [20] Hongyu Song, Rishabh Dev Yadav, Cheng Guo, and Wei Pan. Soranav: Adaptive uav task-centric navigation via zeroshot vlm reasoning. *arXiv preprint arXiv:2510.25191*, 2025. 3

- [21] Yasheerah Yaqoot, Muhammad Ahsan Mustafa, Oleg Sautenkov, Artem Lykov, Valerii Serpiva, and Dzmitry Tsetserukou. Uav-vlrr: Vision-language informed nmpc for rapid response in uav search and rescue. In *IEEE Intelligent Vehicles Symposium*, 2025.
- [22] Tianshun Li, Tianyi Huai, Zhen Li, Yichun Gao, Haoang Li, and Xinhui Zheng. Skyvln: Vision-and-language navigation and nmpc control for uavs in urban environments. *arXiv preprint arXiv:2507.06564*, 2025.
- [23] Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. Trihelper: Zero-shot object navigation with dynamic assistance. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10035–10042. IEEE, 2024.
- [24] Lingfeng Zhang, Hao Wang, Erjia Xiao, Xinyao Zhang, Qiang Zhang, Zixuan Jiang, and Renjing Xu. Multi-floor zero-shot object navigation policy. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6416–6422. IEEE, 2025.
- [25] Lingfeng Zhang, Xiaoshuai Hao, Yingbo Tang, Haoxiang Fu, Xinyu Zheng, Pengwei Wang, Zhongyuan Wang, Wenbo Ding, and Shanghang Zhang. *nava*³: Understanding any instruction, navigating anywhere, finding anything. *arXiv preprint arXiv:2508.04598*, 2025. 3
- [26] Peiran Liu, Qiang Zhang, Daojie Peng, Lingfeng Zhang, Yihao Qin, Hang Zhou, Jun Ma, Renjing Xu, and Yiding Ji. Toponav: Topological graphs as a key enabler for advanced object navigation. *arXiv preprint arXiv:2509.01364*, 2025.
- [27] Lingfeng Zhang, Erjia Xiao, Yuchen Zhang, Haoxiang Fu, Ruibin Hu, Yanbiao Ma, Wenbo Ding, Long Chen, Hangjun Ye, and Xiaoshuai Hao. Team xiaomi ev-ad vla: Caption-guided retrieval system for cross-modal drone navigation-technical report for iros 2025 robosense challenge track 4. *arXiv preprint arXiv:2510.02728*, 2025. 2
- [28] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3
- [29] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 3
- [30] Ronghao Dang, Yuqian Yuan, Yunxuan Mao, Kehan Li, Jiangpin Liu, Zhikai Wang, Xin Li, Fan Wang, and Deli Zhao. Rynnc: Bringing mllms into embodied world. *arXiv preprint arXiv:2508.14160*, 2025. 3
- [31] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025. 3
- [32] Dasong Li, Sizhuo Ma, Hang Hua, Wenjie Li, Jian Wang, Chris Wei Zhou, Fengbin Guan, Xin Li, Zihao Yu, Yiting Lu, et al. Vquala 2025 challenge on engagement prediction for short videos: Methods and results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3391–3401, 2025.
- [33] Xinyu Zheng, Yangfan He, Yuhao Luo, Lingfeng Zhang, Jianhui Wang, Tianyu Shi, and Yun Bai. Railway side slope hazard detection system based on generative models. *IEEE Sensors Journal*, 2025.
- [34] Zeying Gong, Rong Li, Tianshuai Hu, Ronghe Qiu, Lingdong Kong, Lingfeng Zhang, Yiyi Ding, Leying Zhang, and Junwei Liang. Stairway to success: Zero-shot floor-aware object-goal navigation via llm-driven coarse-to-fine exploration. *arXiv preprint arXiv:2505.23019*, 2025.
- [35] Erjia Xiao, Lingfeng Zhang, Yingbo Tang, Hao Cheng, Renjing Xu, Wenbo Ding, Lei Zhou, Long Chen, Hangjun Ye, and Xiaoshuai Hao. Team xiaomi ev-ad vla: Learning to navigate socially through proactive risk perception-technical report for iros 2025 robosense challenge social navigation track. *arXiv preprint arXiv:2510.07871*, 2025.
- [36] Xinyu Zheng, Lingfeng Zhang, Yuhao Luo, and Tiange Wang. A generation-based defect detection system for rail transit infrastructure. *High-speed Railway*, 2025. 2
- [37] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 6
- [38] Qiang Zhang, Gang Han, Jingkai Sun, Wen Zhao, Jiahang Cao, Jiaxu Wang, Hao Cheng, Lingfeng Zhang, Yijie Guo, and Renjing Xu. Lips: Large-scale humanoid robot reinforcement learning with parallel-series structures. *arXiv preprint arXiv:2503.08349*, 2025. 2
- [39] Alessandro Giusti, Jérôme Guzzi, Dan C. Cireşan, Fang-Lin He, Juan P. Rodríguez, Flavio Fontana, Matthias Faessler, Christian Forster, Jürgen Schmidhuber, Gianni Di Caro, Davide Scaramuzza, and Luca M. Gambardella. A machine learning approach to visual perception of forest trails for mobile robots. In *IEEE Robotics and Automation Letters*, pages 661–667, 2016. 3
- [40] Fereshteh Sadeghi and Sergey Levine. Cad2rl: Real single-image flight without a single real image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 3
- [41] Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. *arXiv preprint arXiv:2502.13451*, 2025. 3
- [42] Jianlin Ye, Savvas Papaioannou, and Panayiotis Kolios. Vlm-rrt: Vision language model guided rrt search for autonomous uav navigation. In *International Conference on Unmanned Aircraft Systems*, 2025. 3
- [43] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 2024. 3
- [44] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula,

- Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.
- [45] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024.
- [46] Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.
- [47] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, Yu Qiao, and Jifeng Dai. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, 2024.
- [48] Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, and Bolei Zhou. Embodied scene understanding for vision language models via metavqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [49] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.
- [50] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at Proceedings of the IEEE/CVF International Conference on Computer Vision*’25, 2025.
- [51] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.
- [52] Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, et al. Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning. *arXiv preprint arXiv:2508.19542*, 2025. 3
- [53] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015. 3
- [54] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3
- [55] Sijie Wang, Siqi Li, Yawei Zhang, Shangshu Yu, Shenghai Yuan, Rui She, Quanjiang Guo, JinXuan Zheng, Ong Kang Howe, Leonrich Chandra, et al. Uavscenes: A multi-modal dataset for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28946–28958, 2025. 3
- [56] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5, 7
- [57] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5, 7
- [58] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 6, 7, 8
- [59] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7
- [60] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 7
- [61] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 7
- [62] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 7
- [63] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 7
- [64] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025. 7
- [65] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6